

Capital Bike Share Predictive Model Report

Prepared By: Clark P. Necciai Jr.

November 12, 2023

Contents

1	Executive Summary	2
2	Problem Statement and Approach	2
3	Methodology	2
3.1	Data Preprocessing	2
3.2	Exploratory Data Analysis	3
3.2.1	Univariate Analysis	3
3.3	Correlation	5
3.4	Notable Multivariate Analysis	6
3.4.1	Temperature and Target Relationship	6
3.4.2	Hourly Rental Trends	6
3.5	Feature Reduction	6
3.6	Feature Engineering	7
4	Model Building	7
4.1	Data Partition	7
4.2	Initial Full Model(s)	7
4.2.1	Model Assumptions	7
4.2.2	Model Improvement	7
4.3	Stepwise Selection (Reduced Model)	8
4.3.1	Model Assumptions	8
4.3.2	Test Set Evaluation	8
5	Conclusions and Recommendations	8
6	Appendix	9

1 Executive Summary

Capital Bike Share provides a network of multi-purpose bicycles to the denizens of the Washington D.C. Metropolitan region. We were approached by Capital Bike Share to delve into their hour-by-hour observations across the 2011 and 2012 time frame. Preliminary insights into the dataset provided to us revealed that demand usage for these bicycles can be affected by a variety of noteworthy, influential factors. These can be broken down into two major categories of influence: time and weather.

We examined multi-variable trends and deliberated on patterns which revealed key insights. Based on our visualizations, we concluded rental demand as being dominated by registered riders using bicycles as a primary mode of transportation to and from work. Furthermore, our resulting predictive model confirmed our preliminary analysis, as it was ultimately determined that time-based variables were the most significant factors in accurately predicting bicycle rentals.

2 Problem Statement and Approach

Our primary tasks in this analysis were twofold:

- Identifying the most influential variables relative to their predictive power in determining the total hour-by-hour bicycle rentals and,
- Fitting a multiple regression model predicting the demand for the total hour-by-hour bicycle rentals.

Beginning with an inspection of our dataset and subsequent exploratory data analysis, we aimed to systematically determine those variables which we believed have significant relation to the target variable. Our dataset consisted of 17,379 observations across 17 features. After a well-documented, granular analysis, including univariate and multivariate visualizations, we utilized our findings in a comprehensive modeling process. This process culminated in a multiple linear regression model that not only accurately predicted rental demand, but simultaneously provided an assumption-backed evaluation confirming our models' reliability.

Below we have provided the methodology of our approach, beginning with an inspection and analysis of our data, followed by our modeling selection strategy and diagnostic testing to ensure model generalizability. Finally, we conclude with our recommendations and takeaways.

3 Methodology

3.1 Data Preprocessing

We began our approach with an overview of the integrity of our dataset's structure. We discovered no missing values nor duplicated observations. Neither imputation nor duplicate observation removal was needed. We did, however, note variables which had data types that were non-representative of the underlying values

Variables `dteday`, `season`, `yr`, `mnth`, `hr`, `holiday`, `weekday`, `workingday`, and `weathersit` were all considered to have data types and values which were unclear and in need of re-evaluation. As such, we decided to apply new data types to these variables, effectively categorizing them and applying appropriate labels. The `dteday`, `yr`, `mnth`, `hr`, `weekday`, `weathersit`, and `cnt` were renamed for additional clarity.

3.2 Exploratory Data Analysis

To better understand our variables' distributions, values, and relationships with each other and our target variable, we conducted a thorough exploratory data analysis. Our primary focus is to examine each of our variables' summary statistics and distributions while investigating the underlying values for patterns, inconsistencies, and anomalies which may affect our analysis.

3.2.1 Univariate Analysis

3.2.1.1 Target Variable

3.2.1.1.1 `count_rentals` We are focused primarily on finding relationships significant to and in predicting the target variable, `count_rentals`(previously `cnt`). Each row/observation in our dataset is taken on an hourly basis. Our target is that corresponding hourly count of rentals that have occurred. Our histogram displaying the distribution overlaid with a density line shows us graphically that `count_rentals` is moderately right skewed. There is serious variability in the hour-by-hour rentals, with a minimum of a single rental to a maximum of nearly a thousand rentals per hour, but with 50% of our observations being less than **142** bicycle rentals per hour, marked by the red dashed line.

3.2.1.2 Categorical Variables

3.2.1.2.1 `date` The `date` distribution revealed that not every unique value of date has an expected equivalent number of hourly interval measurements. When considering the **731** distinct date values present in our dataset, the majority of each contained twenty-four or twenty-three of the expected hourly observations. However, fourteen of the dates contained less and in some cases contained a mere single hourly observation, which in total equated to **233** hours worth of missing observations. Capital Bike Shares' hour-to-hour observational system has hourly gaps, which would otherwise provide useful descriptive and predictive analytics.

3.2.1.2.2 `season` As a whole, the `season` distribution was evenly distributed with each season containing approximately within +/- 1% of a quarter of the observations as would be expected.

3.2.1.2.3 `year` The distribution of observations for the two recorded years, 2011 and 2012, were nearly exact at 49.74% and 50.26%, respectively.

3.2.1.2.4 month Across all monthly grouped observations, the distribution of the **month** variable was approximately uniform. However, the month of February appeared unique with it having the fewest number of observations of 1341. Otherwise, this distribution was mostly even.

3.2.1.2.5 hour When observing a sorted hourly distribution from 12:00AM to 11:00PM, we found that the count of each distinct hour were nearly equal. However, a trend in the number of observations can be seen with a decrease beginning at approximately 1:00AM and continuing until 3:00AM when it then increases through 6:00AM.

3.2.1.2.6 day The distribution across the **day** variable were all nearly approximate for each of the seven days of the week.

3.2.1.2.7 weather A vast majority of our **weather** observations were of Type 1, meaning that a majority of our observations were recorded in favorable weather conditions. The second largest proportion of observations were of Type 2 which included cloudy and misty weather. The third largest were of Type 3 which indicates rain/snow and thunderstorm conditions. The smallest proportion were those recorded for Type 4. These observations were of extreme weather conditions, including heavy rain/snow, severe thunderstorms/etc.

3.2.1.3 Continuous Variables Note: Additional summary statistics for each of the continuous variables can be found in the Appendix in Figure 6. Each of our continuous variables were pre-normalized which in turn will help us determine their importance in predicting the target with respect to one another. Additionally, each continuous variables distribution can be found in Figure X.

3.2.1.3.1 temp The distribution of **temp** has been normalized to represent the actual temperature in Celsius.

3.2.1.3.2 atemp The distribution of **atemp** has been normalized and represents the “real-feel” temperature in Celsius. Being that **atemp** is undoubtedly related to influences from the season and weather, our intuition tells us that **atemp** will be an important factor in determining bicycles rentals.

3.2.1.3.3 hum Despite the distribution of **hum** representing humidity being normalized, we see that a majority of the observations contain some degree of humidity. Being that **hum** possibly includes rain, snow, or other precipitation, we might see that it negatively impacts the number of rentals.

3.2.1.3.4 windspeed We have a uniquely zero-inflated, skewed distribution with the **windspeed** variable. Quite a few of our observations are marked as having zero **windspeed**, with a notable gap between zero and the next marked observation. It may be the case that these readings are of truly zero **windspeed** or that the wind was so negligible within this gap of ranges that they were marked as zero.

3.2.1.4 Boolean(True/False) Variables

3.2.1.4.1 holiday Inspection of our `holiday` variable revealed dates which should have been marked as holidays and others which should not have been. We re-labeled these observations based on official [federally recognized holidays](#).

In comparison to our dataset: **2011-01-01, 2011-12-25, 2012-01-01, and 2012-11-11** were incorrectly mislabeled as *not* being holidays. **2011-12-26, 2012-01-02, and 2012-11-12** were mislabeled as **being** holidays. The aforementioned dates were corrected as to maintain consistency with federally recognized holidays. **2011-04-15 and 2012-04-16**, while not being federally recognized holidays, are dates of observance for Emancipation Day in the Washington D.C. Area. Due to the similar holiday-like observance of Emancipation Day in our area of interest, these two dates will be labeled as holidays.

3.2.1.4.2 workingday The distribution of `workingday` shows more observations consisting of days in which people worked than not. This is consistent when considering that no Saturday, Sunday, weekday/end holidays can be working days.

3.3 Correlation

We decided to investigate the extent to which our variables are linearly related to one another with a visualized correlation matrix^[Figure XX]. Here we are concerned with those relationships related to that of our target, `count_rentals` and continuous predictor variables. Here we list the most noteworthy findings:

- Being that `temp` and `atemp` both aim to measure normalized Celsius, it is unsurprising that these variables exhibit a near exact linear relationship with one another. We additionally see they both are moderately, positively related to our target variable to the same degree. Naively including both `temp` and `atemp` in our modeling process would result in multicollinearity, in which we have variables competing to explain our target variable with redundant information. This can cause our estimates our model coefficients to become unstable.

To maintain a robust regression model and avoid multicollinearity, we will choose only one of these variables, `atemp`, to be including in the modeling process.

- Our target is to some degree negatively correlated with `hum`, or the humidity. Being that humidity includes potential precipitation such as rain, mist, snow, and others, it is unsurprising we might find that `count_rentals` decreases as `hum` increases.

Overall, we see that each of our continuous predictor variables exhibits straight-line relationship with our target variable. This can be seen by examining the bottom row of Figure XX as indicated by scatter-plots and the red trend line.

3.4 Notable Multivariate Analysis

3.4.1 Temperature and Target Relationship

We examined the multi-variable relationship between our temperature related variables, `temp`, `atemp`, and `count_rentals`. Based on our graph, we found that unsurprisingly, the highest numbers of rentals occurred in normalized Celsius values we believe to be indicative of favorable conditions.

This plot also revealed twenty-four observations having the exact same `atemp` value of **0.2424**, but `temp` values which were relatively high and variable in relation, ranging from **0.62 to 0.86**. Delving into these observations, we found that all were recorded sequentially, on the exact same day(2012-08-17) and under the same weather conditions(*Type 1*). The changing levels of humidity and wind speed lead us to further believe that the real feel temperature should have likewise varied when compared to other observations of similar values in the dataset.

When considering the evidence of these twenty-four `temp` and `atemp` multivariate as being observations, we will opt to remove these observations. This removal helps to ensure data integrity as we move towards the modeling process.

3.4.2 Hourly Rental Trends

Due to the widespread availability of bicycles as a mode of transportation at all hours of the day, our intuition led us to investigate possible patterns of usage based solely on time. What we uncovered, showed that during working days, there is strong indication that bicycles are being utilized as a primary mode of transportation to and from work. Most notably, it is Capital Bike Shares `registered` rental users that are driving this trend. We see that between all rentals and only registered, the trend is nearly identical[Figure XX, Figure XX]. Neither `casual` rentals nor rentals occurring on Saturday/Sunday rentals exhibit this trend of bicycle usage, adding further evidence to our speculation[Figure XX, Figure XX].

It may be the case that Capital Bike Share bicycles are used as a primary mode of transportation for the majority of `registered` riders during working hours. Before 9:00AM, we note an upward trend of rentals followed by a stark decrease thereafter. We interpret this as individuals arriving at work followed by a sharp increase of usage around 4:00PM/5:00PM when people typically leave work and commute home.

3.5 Feature Reduction

`date` - Our team reasoned that the creation of **731** distinct variables representing `date` could have severe impacts (particularly high VIF and overfitting) on model performance when regressing our target variable. No unique information is provided by `date`. All other information is already found in our `year` and `month`, and `day` variable.

To avoid these drawbacks, we have decided to disregard the `date` variable from the model building process.

3.6 Feature Engineering

Dummy Variable Creation - Categorical variables, such as `season`, `year`, `month`, `day`, `hour`, and `weather`, are better suited for dummy variable creation for use in our multiple linear regression model. This encoding ensures these factor/categorical variables are appropriately interpreted by our model. For each of these variables, one of the categories will not appear in the model summary output, due to it being considered the baseline.

4 Model Building

4.1 Data Partition

With our goals centered on finding the most significant predictors with respect to determining the total number of hour-by-hour rentals, we decided to partition our data 80% towards training the model and 20% for testing model performance. Allowing a significant majority of our overall dataset to go towards the model fitting process enables us to get better coefficient estimates as long as we have clean, non-noisy data.

4.2 Initial Full Model(s)

4.2.1 Model Assumptions

4.2.1.1 Normality

4.2.1.2 Homoscedasticity

4.2.1.3 Multicollinearity

4.2.2 Model Improvement

4.2.2.1 Model Outliers/Influential Observations Each of our three `weather_type_4` observations had significantly large leverage on our model fitting. Because of this we decided it would be best to remove these. Because of this, the `weather_type_4` variable consequently was removed from the training data set.

4.2.2.2 Target Variable Transformation [REWRITE]

during this analysis is to fit multiple linear regression models that carry certain assumptions which must be satisfied. These mainly include our response variable being linearly related to the predictors(linearity) and equality of variances in the residuals(homoscedasticity). With their current skewed distributions, these two assumptions would undoubtedly be violated post-model building.

In order to satisfy the assumptions of linear regression, a transformation on each of the target variables becomes necessary. We considered multiple transformations to approach normality, including, the Box-cox, logarithmic, square-root, cube-root, and fourth-root transformations.

For each of our target variables, the transformation which approximated normality (closest skewness value to 0) the closest was the **cube-root** transformation. Neither the logarithmic and Box-cox transformations could not be applied to **casual** and **registered** due to the presence of zero(0) values which are incompatible with the transformations.

4.2.2.3 Multicollinearity (VIF)

4.3 Stepwise Selection (Reduced Model)

4.3.1 Model Assumptions

4.3.2 Test Set Evaluation

Final Model Equation

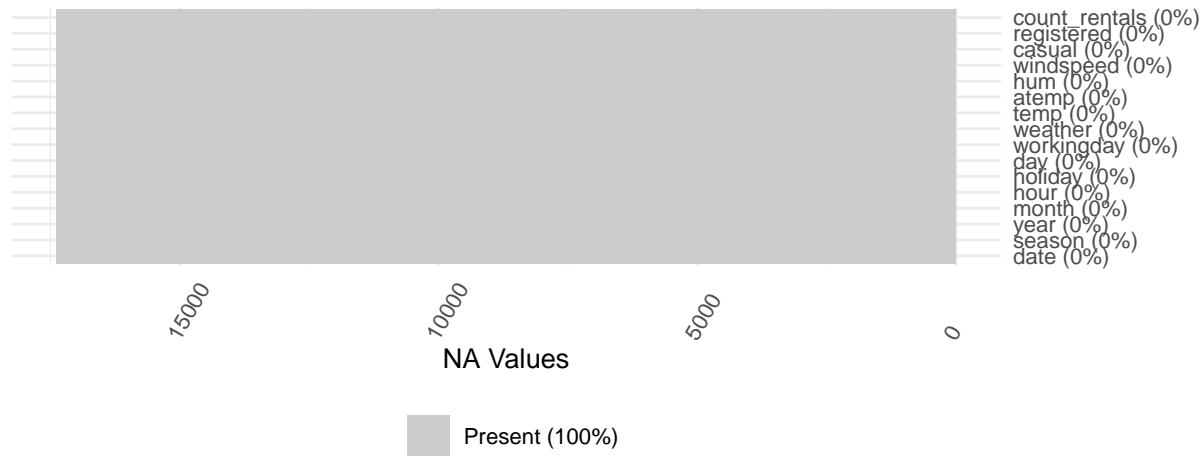
$$\begin{aligned} \text{cube_root_total} = & \beta_0 + \beta_1 \text{holiday} + \beta_2 \text{atemp} + \beta_3 \text{hum} + \beta_4 \text{season_Spring} + \\ & \beta_5 \text{season_Fall} + \beta_6 \text{year_2012} + \beta_7 \text{month_May} + \beta_8 \text{month_Aug} + \beta_9 \text{month_Sep} \\ & + \beta_{10} \text{day_Friday} + \beta_{11} \text{day_Saturday} + \beta_{12} \text{hour_1:00PM} + \beta_{13} \text{hour_10:00AM} + \\ & \beta_{14} \text{hour_10:00PM} + \beta_{15} \text{hour_11:00AM} + \beta_{16} \text{hour_11:00PM} + \beta_{17} \text{hour_12:00AM} + \\ & \beta_{18} \text{hour_12:00PM} + \beta_{19} \text{hour_2:00AM} + \beta_{20} \text{hour_2:00PM} + \beta_{21} \text{hour_3:00AM} + \beta_{22} \text{hour_3:00PM} \\ & + \beta_{23} \text{hour_4:00AM} + \beta_{24} \text{hour_4:00PM} + \beta_{25} \text{hour_5:00AM} + \beta_{26} \text{hour_5:00PM} + \\ & \beta_{27} \text{hour_6:00AM} + \beta_{28} \text{hour_6:00PM} + \beta_{29} \text{hour_7:00AM} + \beta_{30} \text{hour_7:00PM} + \beta_{31} \text{hour_8:00AM} \\ & + \beta_{32} \text{hour_8:00PM} + \beta_{33} \text{hour_9:00AM} + \beta_{34} \text{hour_9:00PM} + \beta_{35} \text{weather_Type_2} + \\ & \beta_{36} \text{weather_Type_3} + \epsilon \end{aligned}$$

5 Conclusions and Recommendations

Capital Bike Share Needs to Capitalize on the trend of registered riders using it as a primary mode of transportation and set of locations at major areas of work and near residential areas where people can get easy access to these bikes. This will increase the number of registered riders.

6 Appendix

```
# Visualize for Missing Data
vis_miss(CBS, sort_miss = T) + labs(y = "NA Values") + theme(axis.text.x.bottom = element_text(vjust = 0)) + coord_flip()
```



```
# Visualize for Missing Data
vis_dat(CBS) + theme(axis.title.y = element_blank(), axis.text.x.bottom = element_blank()) +
  labs(title = "Data Types of Capital Bike Share Dataset") +
  scale_fill_brewer(palette = 2)
```

Data Types of Capital Bike Share Dataset

